# Statistical models for the development of psychological and educational tests

Sijtsma, K.; Emons, W.H.M.

*Published in:*
Handbook of probability

*Publication date:*
2008

# 16

# STATISTICAL MODELS FOR THE DEVELOPMENT OF PSYCHOLOGICAL AND EDUCATIONAL TESTS

KLAAS SIJTSMA AND WILCO H. M. EMONS

## INTRODUCTION

The general idea behind modern measurement in the social and behavioral sciences is that human behavior is driven by a limited number of traits, attitudes, opinions, skills, and abilities. Each of these attributes serves as an explanation of the cohesion in certain sets of observable behaviors and constitutes what the researcher is really interested in. Examples are as follows. Clinical and personality psychologists may be interested in traits such as introversion and anxiety, sociologists in attitudes toward euthanasia or religiosity, and developmental psychologists in the Piagetian developmental abilities of conservation and transitive reasoning. Many interesting applications can be found in other disciplines such as education (e.g., knowledge of disciplines taught at school), marketing research (e.g., service quality of medical facilities), political science (e.g., opinions about government policy), and social medicine (e.g., quality of life after surgery). The philosophical status of these attributes has been debated in several sources (e.g., Borsboom, 2005; Michell, 1990). Here, we will simply take them for granted as organizing principles behind cohesive sets of observable behaviors.

This takes us to the way these attributes are measured. Because these attributes are latent, conclusions about them have to be inferred from sets of cohesive behaviors that are assumed to be driven by these attributes. Thus, in practice, evidence on the unifying cause comes from what is believed to be its effect—that is, the data collected on the set of observable behaviors that are assumed to be typical of this cause. This is done as follows. A set of $J$ stimuli—questions, statements, tasks: *items*, for short—is presented to a representative sample of $N$ respondents from the population of interest, and each respondent provides responses to each item. Responses can be choices from a set of answers, as in selected-response items (e.g., multiple-choice items) for measuring knowledge of national history, or sentences reflecting the answer to a question, as in constructed-response items, ratings on an ordered scale for each attitude statement in a set, or verbal accounts of the process that lead to the solution of transitive reasoning problems.

Tests can have different appearances. For example, the respondent may react to a paper-and-pencil test and encircle response options, write down answers, rate statements, or manipulate real

objects, such as a pen through a maze as in intelligence testing. In computerized testing, similar actions may be performed by pressing keys on a keyboard, moving a mouse, or touching a screen. Likewise, surveys including sets of items for measuring attitudes and opinions need not only be verbal (i.e., as in a street or a telephone interview) or in writing (e.g., as in mail surveys) but may also be administered through the Internet, which in principle enhances their possibilities comparable with computerized testing.

The qualitative responses to items—choices from a number of precoded options or written sentences, ratings on discrete ordinal scales, verbal explanations—are coded next as integers, following the principle that the more evidence a response gives of a higher level of, in these examples, knowledge, attitude, and ability, the higher the item score. Obviously, whether this coding is meaningful depends on the degree to which the items adequately reflect relevant aspects of the attribute of interest. If the theory or the operationalization of this attribute is primitive, or even wrong, responses may have a muddled relationship to the attribute and responses to different items may exhibit little cohesion. Thus, the use of a sound theory and a meaningful operationalization into a set of items are prerequisites for the production of a set of cohesive quantitative item scores that form the basis for the construction of a measurement instrument—a *test*, for short.

Numerous statistical models have been proposed for analyzing the item scores produced by $N$ respondents who reacted to $J$ items (Boomsma, Van Duijn, & Snijders, 2001; Van der Linden & Hambleton, 1997). The application of such models produces information on the following:

- *Dimensionality of the data*—that is, the number of mathematical dimensions needed to explain the data structure. The relations among these dimensions are described in a probabilistic model and are often taken as evidence of one or more explanatory attributes. This may enlighten the meaning of measurement and may or may not confirm the researcher's expectations. From a practical angle, one dimension supports the use of one "measurement rod" or scale for the attribute of interest, and multiple dimensions may call for several scales.
- *Quality of individual items*, such as an item-difficulty parameter, which indicates the ability level required for solving the problem with average probability, and an item-discrimination parameter, which indicates how well the item separates lower ability levels from higher levels. Items that are too easy or too difficult and items that discriminate weakly may be rejected from the final test because they are not properly tuned to the group to be measured.
- *Quality of the whole test*, such as the accuracy of measurement that is possible with a set of items of good quality that together constitute a scale. This accuracy may be expressed in one summary statistic, known as reliability, or as a function of the scale indicating how accurately the test measures at different scale levels. Quality is also expressed as scale validity, indicating the degree to which test performance is driven by the attribute(s) of interest and the degree to which performance on individual items is driven by these attributes. Validity takes the form of a series of results from research rather than a single index or function.

After a scale has been constructed on the basis of information on dimensionality and item and test quality, measurement values for individuals locating them on the scale are determined. These measurement values express the individual's attribute level and can be used to classify the individual for entry or nonentry in a course, for receiving or not receiving therapeutic treatment—either psychological or medical—and for admittance to or rejection from a job. Each of these uses of measurements emphasizes the need for reliable and valid instruments.

The statistical models referred to are united in the family of *item response theory* (IRT) models. The purpose of this chapter is to discuss a few well-known and regularly used models that are representative of the IRT family. Four of these IRT models are used to analyze data from an

arithmetic test. It is explained how these models can be used to construct tests and also how they are complementary to one another. Finally, we discuss other possibilities offered by IRT for data analysis and the construction of scales for the measurement of attributes.

## Assumptions of Item Response Theory

We assume that a test or a questionnaire consists of $J$ items, which are meant to measure the latent attribute(s) of interest. The scores on items are modeled by random variables, $X_j$, indexed $j = 1, \ldots, J$, and are usually integer valued: $X_j = x_j$, with $x_j = 0, 1$, for example, expressing incorrect or correct responding, and $x_j = 0, \ldots, m$, for example, expressing the degree to which someone agrees with an attitude statement. These are the most frequently used possibilities, referred to as dichotomous and polytomous scoring, respectively. The latent attribute is often called the latent trait, where the word *trait* is assumed to also capture personality traits, attitudes, opinions, skills, and abilities, but a neutral term such as *latent variable* would probably fit in better with mainstream statistics. Latent variables are denoted $\theta_q$, with $q = 1, \ldots, Q$, and collected in vector $\boldsymbol{\theta}$.

Three classes of assumptions are relevant for IRT models. The first class of assumptions describes the relationship between the probability of a particular score on item $j$ and the latent variables, denoted $P(X_j = x_j|\boldsymbol{\theta})$. This is the response function. For dichotomously scored items, it is known as the item response function (IRF), $P(X_j = 1|\boldsymbol{\theta}) \equiv P_j(\boldsymbol{\theta})$. Most IRT models assume that the IRF is monotone nondecreasing in $\boldsymbol{\theta}$, coordinate-wise in each element $\theta_q$, $q = 1, \ldots, Q$. This is the monotonicity (M) assumption that says that the probability of, for example, a correct response does not decrease—that is, remains constant or increases—when either one of the $\theta$s increases while the others are kept constant. If one latent variable, say alienation, drives item responses (and thus $\boldsymbol{\theta} = \theta$), then Assumption M says that the probability of saying "Yes" to the question whether one avoids neighborhood festivities does not decrease—often increases—with higher values of $\theta$.

For polytomously scored items, several possibilities for defining response probabilities exist (e.g., Mellenbergh, 1995). One such possibility is $P(X_j \geqslant x_j|\theta)$, with $x_j = 1, \ldots, m$, which is the item step response function (ISRF). Assumption M says that $P(X_j \geqslant x_j|\theta)$ is nondecreasing in $\theta$. For example, a respondent rates on a 5-point scale *to what degree* he or she avoids neighborhood festivities; and Assumption M says that the probability of rating at least the $(x_j + 1)$st category—that is, obtaining at least score $x_j$—does not decrease when level of alienation increases.

The second class of assumptions describes the relationships between the items. Specifically, conditioning on $\boldsymbol{\theta}$ simplifies the joint conditional distribution of the $J$ item scores, collected in vector $\mathbf{X} = (X_1, \ldots, X_J)$ with realization $\mathbf{x}$, into the product of marginal conditional distributions, such that

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(X_j = x_j|\boldsymbol{\theta}). \qquad (16.1)$$

This is the assumption of local independence (LI), in statistics better known as conditional independence. Equation (16.1) implies that for two items $j$ and $k$,

$$\mathrm{Cov}(X_j, X_k|\boldsymbol{\theta}) = 0, \qquad j,k = 1, \ldots, J; \quad j < k, \qquad (16.2)$$

but reversely, LI is not implied by this set of zero covariances. Thus, LI represents a stronger independence property than that represented by the set of $\frac{1}{2}J(J-1)$ conditional covariances in (16.2). Consequently, (16.2) is known as weak local independence (WLI) (Stout, 2002) or, using a more general terminology, conditional uncorrelatedness. Obviously, LI and WLI only hold when $\boldsymbol{\theta}$ contains all $Q$ latent variables relevant for measurement, and failure of these properties in real data is an indication that the dimensionality of the data is different from what the researcher expected. Several procedures have been proposed that explore the data for dimensionality in an effort to approach (16.2) (Stout et al., 1996).

The third set of assumptions refers to the number of latent variables. Typical of psychological

measurement is the requirement that the test measure one latent variable. This renders measurements to unambiguously reflect one "thing" at a time and not a mixture, just as one wants the scale of a thermometer to reflect only temperature and not a mixture of temperature, air pressure, humidity, and wind velocity. Thus, the majority of IRT models assume that $\boldsymbol{\theta} = \theta$ and thus $Q = 1$ (for an overview, see Van der Linden & Hambleton, 1997). This is Assumption $D = 1$. This simplifying assumption is somewhat at odds with psychological reality, whereby responses to items are usually driven by multiple psychological properties (an arithmetic item requires not only arithmetic ability but also reading skills, verbal comprehension, and sometimes also spatial orientation), so that unidimensionality is an ideal and multidimensional models are more realistic (thus assuming $D \geqslant 2$). Nevertheless, unidimensional IRT models are often seen as reasonable approximations to the real dimensionality, which may be defendable when one dominant property drives item responses and the influence of others is minor or may be ignored.

## Special Cases of (M, LI, $D = 1$) Models

In this section, several well-known and much used IRT models are discussed. The most important distinctions are between nonparametric and parametric models and between models for dichotomous and polytomous item scores.

### Monotone Homogeneity Model for Dichotomous Items

#### *Model Formulation*

Mokken (1971) introduced the monotone homogeneity model (MHM) for dichotomously scored items. The MHM is defined by the assumptions of M, LI, and $D = 1$. This model is important in practice because it implies that individuals are measured on an ordinal scale. To see this, define the observable total score

$$X_+ = \sum_{j=1}^{J} X_j, \tag{16.3}$$

and note that for two individuals, $v$ and $w$, with total scores $x_{+v} < x_{+w}$, the MHM implies for each value $t$ of $\theta$ that

$$P(\theta > t | X_+ = x_{+v}) \leqslant P(\theta > t | X_+ = x_{+w}) \tag{16.4}$$

(Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). Equation (16.4) is known as stochastic ordering of the latent variable by the total score (SOL). SOL implies that for expected values ($E$)

$$E(\theta | X_+ = x_{+v}) \leqslant E(\theta | X_+ = x_{+w}). \tag{16.5}$$

SOL means that the observable total score $X_+$ orders individuals on the scale of latent variable $\theta$; thus, a fitting MHM implies an ordinal scale for person measurement.

The fit of the MHM can be investigated in two steps. First, the dimensionality of an item set is investigated, and second, the monotonicity of the IRFs is investigated.

#### *Mokken Scaling and Dimensionality Investigation*

Mokken (1971, chap. 5) and Sijtsma and Molenaar (2002, chap. 5) proposed an *exploratory* item selection procedure that combines the investigation of the dimensionality of the data with an evaluation of the quality of the items found to assess the same dimension. This method selects items into clusters on the basis of the strength of their relationships with the latent variables such that each cluster measures a different $\theta$. Items that predominantly measure a $\theta$ that is not shared by any of the other items are declared unscalable.

Strength of relationship is indexed by means of the item scalability coefficient $H_j$, which is defined as follows. Let $\text{Cov}(X_j, X_k)$ denote the covariance between item scores $X_j$ and $X_k$ and $\text{Cov}(X_j, X_k)_{\max}$ the maximum possible covariance given fixed marginals of the $2 \times 2$ frequency table of bivariate counts; then, $H_j$ is defined as

$$H_j = \frac{\sum_{k \neq j} \text{Cov}(X_j, X_k)}{\sum_{k \neq j} \text{Cov}(X_j, X_k)_{\max}}, \qquad j = 1, \ldots, J. \tag{16.6}$$

For a set of $J$ items evaluated as one test, coefficient $H$ is defined as

$$H = \frac{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \mathrm{Cov}(X_j, X_k)}{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \mathrm{Cov}(X_j, X_k)_{\max}} \tag{16.7}$$

and is seen to be a positively weighted average of the $J$ item coefficients, $H_j$ $(j = 1, \ldots, J)$ (Mokken, 1971, pp. 148–153),

$$H = \frac{\sum_{j=1}^{J} \sum_{k \neq j} \mathrm{Cov}(X_j, X_k)_{\max} H_j}{\sum_{j=1}^{J} \sum_{k \neq j} \mathrm{Cov}(X_j, X_k)_{\max}}, \tag{16.8}$$

such that $H$ is bounded by

$$\min(H_j) \leqslant H \leqslant \max(H_j), \qquad j = 1, \ldots, J. \tag{16.9}$$

Given the interpretation of $H_j$, coefficient $H$ indexes the average strength of relationship of the $J$ items with the latent variable $\theta$. The stronger this relationship, the better—more accurately—the test separates relatively low $\theta$s from relatively high $\theta$s (Mokken, Lewis, & Sijtsma, 1986). Thus, if the MHM holds, a high $H$ indicates accurate person ordering by means of $X_+$.

For the class of (M, LI, $D = 1$) models—MHM and special cases that we shall encounter shortly—it can be shown that

$$\mathrm{Cov}(X_j, X_k) \geqslant 0, \qquad \text{for all } (j,k), \quad j \neq k. \tag{16.10}$$

Because IRFs are nonlinear, other association measures may be in order; see Holland and Rosenbaum (1986) for suggestions and also a more general positive covariance condition known as conditional association of which (16.10) is a special case. Using the positive-sign property of (16.10), it follows that

$$0 \leqslant H \leqslant 1 \tag{16.11}$$

and, similarly, for the item coefficients, that

$$0 \leqslant H_j \leqslant 1, \qquad j = 1, \ldots, J. \tag{16.12}$$

Thus, positive values of $H$ and $H_j$ are necessary conditions for the MHM model to hold; hence, negative values are in conflict with the model.

Mokken (1971, p. 184) defined a scale as a set of items for which, denoting correlation by $\rho$ and given a suitably chosen constant $c$,

1. $\rho_{jk} > 0$ for all item pairs $(j,k)$, $j \neq k$ and
2. $H_j \geqslant c > 0$ for all items $j$.

Positive correlations and positive $H_j$s both are implied by (16.10), and requiring a positive lower bound $c$ means that only those items are admitted in the scale that have a positive relationship with $\theta$, the strength of which is controlled by the magnitude of $c$.

*Exploratory* item analysis focuses on selecting items in the same subset that have high $H_j$s relative to one another and low $H_j$s relative to items that are in another subset. High $H_j$s are due to the same common $\theta$ assessed by the items in the same subset, and low $H_j$s express a weak relationship with the $\theta$ assessed by the items in the other subset.

The algorithm that does the item selection is a bottom-up procedure that selects items one by one, starting with the pair out of $\frac{1}{2}J(J-1)$ candidate pairs that has the highest, significantly positive $H_{jk}$ value (this is $H$ for two items). In each of the next selection steps, from the unselected items an item is selected such that (1) it has a positive correlation with the items already selected and (2) its $H_j$ relative to the items already selected is significantly greater than 0 and also $H_j > c$, and if more items satisfy Conditions (1) and (2), from this set the item is selected that (3) together with the items already selected in previous steps of the algorithm produces the greatest common $H$. This results in a subset of items that predominantly measure the same $\theta$, while a high value of $H$ in (16.8) guarantees accuracy of ordinal person measurement in the sense of SOL (16.4) that is controlled by the choice of lower bound $c$.

If the data are unidimensional, in principle all items fit in the same cluster. However, if the items have different $H_j$s, which is the common situation in practice, higher $c$ values may cause more items to remain unselected. This is not because they do not assess $\theta$ but because they do so more weakly than $c$ allows. The researcher should decide what he or she considers a desirable outcome and may take considerations into account such

as the degree to which only few items can adequately cover the attribute well.

For multidimensional data—say each subset of items assesses a particular $\theta$, and different subsets assess different $\theta$s, to keep things simple—the typical sequence of outcomes is that, first, low $c$ values (near 0) lead to the selection of (nearly) all items in one cluster and, second, higher $c$ values result in the clustering that reflects true dimensionality. Hemker, Sijtsma, and Molenaar (1995) recommended running the cluster algorithm for different $c$ values, starting at 0, using increments of 0.05, and stopping at 0.6. The data section in this chapter will offer an example.

*Confirmatory* item analysis evaluates a set of $J$ items as a given scale. This situation is relevant when the researcher is interested in testing the hypothesis that a newly constructed test represents a scale. Also, he or she may consider one or more items in an existing test to have become archaic—for example, due to the use of old-fashioned words—and have them replaced by others or the instrument may be investigated for use in another population. In each of these cases, the researcher takes the $J$-item test as given and estimates its $H$ and $H_j$ coefficients to assess test score and item quality, respectively.

Stout et al. (1996) have proposed a method for dimensionality investigation that searches for the partitioning of the item set that approximates WLI (16.2) as well as possible but without taking item quality into consideration (Van Abswoude, Van der Ark, & Sijtsma, 2004) as Mokken's method typically does. These and other methods have been compared by Van Abswoude et al. (2004).

### *Monotonicity Investigation*

In real data, the relationship between item and latent variable may be monotone, as the MHM assumes, but it is regularly found that for some items in the test the relationship is either monotone by approximation—the empirical curve tends to increase but shows several small local decreases—or sometimes even distinctly nonmonotone. Mokken's method selects items having $H_j$s of at least $c$ in subsets, which ascertains IRFs that show at least a tendency to increase in $\theta$, just as a regression curve with a positive regression coefficient does. The higher the value of $c$, the stronger this tendency and, roughly, the smaller the chances that local decreases are such that the curve can no longer be evaluated to be approximately monotone. Thus, for most $c$ values, within selected item subsets the additional investigation of Assumption M is useful, and this is true a fortiori the smaller $c$ is. Assumption M is investigated as follows.

Define a total score without Item $j$, called a restscore and denoted $R_{(-j)}$, as

$$R_{(-j)} = \sum_{k \neq j} X_k. \tag{16.13}$$

Like $X_+$, restscore $R_{(-j)}$ estimates person ordering on $\theta$, which is justified by the same stochastic ordering results (16.4). The MHM implies manifest monotonicity (MM) (Junker, 1993),

$$P[X_j = 1 | R_{(-j)} = r] \text{ nondecreasing in } r = 0, \ldots, J-1. \tag{16.14}$$

Junker and Sijtsma (2000) showed that an MM result as in (16.14) is not obtained when $R_{(-j)}$ is replaced by $X_+$. MM can be used to estimate the IRF by means of nonparametric regression. One straightforward possibility is to estimate for each value of $r$ the proportion of the population that have Item $j$ correct, plotting these proportions as a function of $r$ and then checking visually for MM and testing local decreases for significance by means of a normal approximation to the binomial test (Molenaar & Sijtsma, 2000). This approach yields a limited number of at most $J$ discrete points of the IRF. Ramsay (1991) proposed a kernel smoothing approach to obtain a continuous estimate of the IRF. Karabatsos and Sheu (2004) discuss a Bayesian approach to evaluating Assumption M.

### *Computer Software*

In our data example, we used the program MSP (Molenaar & Sijtsma, 2000) to estimate the $H$ and $H_j$ coefficients and select items into clusters for different $c$ values, and also to estimate discrete versions of the IRFs. The program TestGraf98 (Ramsay, 2000) was used to estimate continuous versions of the IRFs.

## Monotone Homogeneity Model for Polytomous Items

Molenaar (1997) generalized the MHM to polytomous item scores by redefining Assumption M for conditional probability $P(X_j \geqslant x_j|\theta)$, for $x_j = 1,\ldots,m$. Obvious as this choice may seem, it has been found to have many far-reaching consequences at the theoretical level, which show that the generalization of dichotomous-item models to polytomous-item models may be problematic. Here are two consequences.

First, Hemker et al. (1997) found that the SOL property does not hold for the polytomous-item MHM or for most other IRT models for ordered polytomous items. Van der Ark (2005) established SOL in many data sets by producing a wealth of robustness results for several, much used polytomous-item models, thus demonstrating convincingly that ordinal measurement properties could be maintained at the practical level.

Second, $P[X_j \geqslant x_j|R_{(-j)}]$ has been shown not to be monotone in general, thus losing MM (Junker & Sijtsma, 2000). Specifically, a nondecreasing observable curve, $P[X_j \geqslant x_j|R_{(-j)}]$, is neither a necessary nor a sufficient condition for Assumption M, but much practical experience with simulated data suggests that such monotone curves tend to be supportive of Assumption M. Software for estimating these curves is available (Molenaar & Sijtsma, 2000; Ramsay, 2000).

Fortunately, as concerns dimensionality analysis, Mokken's item selection method has been generalized successfully by defining coefficients $H_j$ and $H$ for polytomous items, maintaining the properties in (16.8), (16.9), (16.11), and (16.12). The program MSP can be used here as well.

## The Three- and Two-Parameter Logistic Models

### *Model Formulation*

The three-parameter logistic model (3PLM) (Birnbaum, 1968) is an (M, LI, $D = 1$) model that specializes Assumption M to a logistic IRF with three item parameters,

$$P_j(\theta) = \gamma_j + \frac{(1-\gamma_j)\exp[\alpha_j(\theta-\delta_j)]}{1+\exp[\alpha_j(\theta-\delta_j)]}, \quad 0<\gamma_j<1, \quad \alpha_j>0. \tag{16.15}$$

In (16.15), parameter $\gamma_j$ is the lower asymptote for $\theta \to -\infty$, parameter $\delta_j$ is the location or difficulty parameter, and parameter $\alpha_j$ is the steepest slope or discrimination parameter of the IRF, evaluated at the point with coordinates $(\delta_j, \frac{1+\gamma_j}{2})$. After it has been established whether the data are unidimensional and the smooth S-shaped IRFs in (16.15) fit the data, these item parameters are estimated and then summarize three important aspects of the item: Positive (i.e., nonzero) $\gamma_j$ indicates that people with low $\theta$s have a nontrivial probability of giving the correct answer, as in multiple-choice items; $\delta_j$ indicates the degree to which the item is difficult for the population of interest; and $\alpha_j$ indicates the degree to which the item separates $\theta$s that are low compared with $\delta$ from $\theta$s that are high compared with $\delta$.

The two-parameter logistic model (2PLM) (Birnbaum, 1968) specializes Assumption M of the 3PLM by assuming that $\gamma_j = 0$, $j = 1,\ldots,J$, resulting in

$$P_j(\theta) = \frac{\exp[\alpha_j(\theta-\delta_j)]}{1+\exp[\alpha_j(\theta-\delta_j)]}, \quad \alpha_j>0. \tag{16.16}$$

The interpretation of the remaining two item parameters is the same as in the 3PLM.

### *Estimating the 3PLM and the 2PLM*

The estimation of the item parameters and the latent variable is straightforward. Let $\mathbf{X}_{N\times J}$ be the data matrix produced by a sample of $N$ individuals, indexed by $v$, who responded to $J$ items. Also, let $\boldsymbol{\theta}_N = (\theta_1,\ldots,\theta_N)$ and $\boldsymbol{\omega} = (\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{\alpha}) = (\gamma_1,\ldots,\gamma_J,\delta_1,\ldots,\delta_J,\alpha_1,\ldots,\alpha_J)$; then, assuming independent and identically distributed (iid)-sampled individuals and LI (16.1), the likelihood of the data can be written as

$$\begin{aligned} L(\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}|\boldsymbol{\theta}_N,\boldsymbol{\omega}) &= \prod_{v=1}^{N}\prod_{j=1}^{J} P_j(\theta_v)^{x_{vj}}[1-P_j(\theta_v)]^{1-x_{vj}}, \end{aligned} \tag{16.17}$$

with (16.15) inserted for $P_j(\theta_v)$ or (16.16) inserted for $P_j(\theta_v)$ and $\omega^* = (\delta, \alpha)$ replacing $\omega$. Several methods have been proposed for estimating the parameters taking this likelihood as a starting point. The oldest method is joint maximum likelihood (JML) estimation, which maximizes the likelihood in (16.17) simultaneously for all parameters in $\theta$ and $\omega$. However, JML has been shown to fail because in the presence of $N$ incidental parameters in $\theta$, the structural parameters in $\omega$ are estimated inconsistently (Neyman & Scott, 1948). Marginal maximum likelihood (MML) estimation of the item parameters does not suffer from this problem and yields consistent estimates for the item parameters in $\omega$ as the number $N$ of respondents grows. We will briefly review the much used MML method.

Define the problem as follows. Let $f(\theta)$ denote the probability density of $\theta$ with parameters collected in $\tau$; then, the marginal likelihood is

$$P[\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}|\omega,\tau] = \prod_{v=1}^{N}\int_{\theta}\prod_{j=1}^{J} P_j(\theta)^{x_{vj}}[1-P_j(\theta)]^{1-x_{vj}} f(\theta)d\theta. \tag{16.18}$$

The integral gives the marginal probability of the item-score vector of person $v$, $\mathbf{x}_v$, which can be denoted by $P(\mathbf{x}_v|\omega,\tau)$, so that we may define

$$P(\mathbf{x}_v|\omega,\tau) = \int_{\theta}\prod_{j=1}^{J} P_j(\theta)^{x_{vj}}[1-P_j(\theta)]^{1-x_{vj}} f(\theta)d\theta \tag{16.19}$$

and write the marginal likelihood as

$$P[\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}|\omega,\tau] = \prod_{v=1}^{N} P(\mathbf{x}_v|\omega,\tau). \tag{16.20}$$

Often, the normal density is chosen for $f(\theta)$, with parameters $\tau = (\mu, \sigma^2)$.

The probability on the left in (16.20) is a function of $3J$ parameters in $\omega$, and these, as well as those in $\tau$, can be estimated by MML (see Bock & Lieberman, 1970, for details;, also see Baker & Kim, 2004, chap. 6). Estimation of $\theta_N$ then follows from evaluating the posterior distribution of each $\theta_v$, denoted $P(\theta_v|\mathbf{x}_v;\omega,\tau)$ and computed by means of Bayes's theorem,

$$P(\theta_v|\mathbf{x}_v;\omega,\tau) = \frac{P(\mathbf{x}_v|\theta_v;\omega)f(\theta_v|\tau)}{P(\mathbf{x}_v|\omega,\tau)}. \tag{16.21}$$

In (16.21), $f(\theta_v|\tau)$ serves as the prior density of $\theta_v$ and is assumed to be the same for each $\theta$ value. The probability of person $v$'s data in $\mathbf{x}_v$ is weighted by the density of each $\theta$ from the prior, and given the marginal likelihood in the denominator, which is independent of $\theta$, this results in the posterior of $\theta_v$. The mean of this posterior often is taken as the estimate of $\theta_v$ (e.g., Bock & Mislevy, 1982).

*Measurement Accuracy*

Fisher's information function expresses the measurement quality of one or more items relative to the latent variable. Let $P_j'(\theta)$ be the first derivative of the IRF with respect to $\theta$. Then, for $Q_j(\theta) \equiv 1 - P_j(\theta)$, Fisher's information function for item $j$, denoted $I_j(\theta)$, is

$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{P_j(\theta)Q_j(\theta)}, \tag{16.22}$$

and given LI, Fisher's information function for the $J$-item test equals

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta). \tag{16.23}$$

Insertion of (16.15) and (16.16) in (16.23) gives the test information functions for the 3PLM and the 2PLM, respectively.

The test information function provides the statistical information in the $J$ items together for estimating $\theta$, and $I(\theta)^{-1/2}$ gives a lower bound on the standard error for estimated $\theta$, which is achieved asymptotically for maximum likelihood (ML) estimation as $J \to \infty$. Suppose one wants a test to measure accurately at a cutoff score denoted $\theta_0$, then test information, $I(\theta_0)$, should have a value high enough to result in a standard error that is sufficiently small for the test application envisaged. This can be accomplished by selecting items that contribute relatively large $I_j(\theta_0)$ values to $I(\theta_0)$. Equation (16.22) shows that the IRFs of these items have relatively steep slopes at $\theta_0$. Van der Linden (2005) discusses many examples of test construction based on this item selection principle.

*Fitting the 3PLM and the 2PLM*

For short tests ($J < 20$), the standardized posterior residuals, also known as root-mean square deviates (RMSDs), are evaluated (Zimowski, Muraki, Mislevy, & Bock, 1996). The RMSD is based on the standardized differences between the posterior probability of a correct response at selected values of $\theta$ and the expected probability at those $\theta$ values. RMSD $> 2.0$ indicates item misfit.

*Computer Software*

The program BILOG-MG (Zimowski et al., 1996) was used to estimate both the 3PLM and the 2PLM and evaluate their fit. Parameters were estimated using MML, and the RMSD was used to assess item fit.

## The Graded Response Model

To our knowledge, a feasible generalization of the 3PLM to polytomous items does not exist to date. The most direct generalization of the 2PLM to polytomous items is the graded response model (GRM) (Samejima, 1997). The GRM is an (M, LI, $D = 1$) model that specializes the ISRF, $P(X_j \geqslant x_j|\theta)$, as

$$P(X_j \geqslant x_j|\theta) = \frac{\exp[\alpha_j(\theta - \delta_{jx_j})]}{1+\exp[\alpha_j(\theta - \delta_{jx_j})]}, \quad x_j > 0, \quad \alpha_j > 0. \tag{16.24}$$

Note that this response function is equivalent to that of the 2PLM but that the difference lies in the item score that is modeled: Polytomous $X_j \geqslant x_j$ in the GRM and binary $X_j = 1$ in the 2PLM, and that they coincide when $m = 1$. The GRM has been characterized as a cumulative probability model (Hemker et al., 1997; Mellenbergh, 1995). Such models are sometimes associated with data stemming from a respondent's global assessment of the rating scale and the consecutive choice of a response option from all available options. Baker and Kim (2004, chap. 8) discuss ML estimation of the item and person parameters of the GRM, and Samejima (1997) discusses goodness-of-fit methods. The program MULTILOG (Thissen, Chen, & Bock, 2003) can be used to estimate parameters and evaluate the fit of the GRM to the data.

## The Rasch Model

*Model Formulation*

The Rasch (1960) model, also known as the one-parameter logistic model (1PLM), is obtained from the 2PLM by setting $\alpha_j = 1$, for $j = 1, \ldots, J$, which results in

$$P_j(\theta) = \frac{\exp(\theta - \delta_j)}{1+\exp(\theta - \delta_j)}. \tag{16.25}$$

Thus, within the same test, items are assumed to separate $\theta$s equally well at different item difficulty levels. Item parameters can be estimated by means of MML (Thissen, 1982), and assuming these estimates to be the true values, person parameters can be estimated by means of ML. Because the Rasch model is a member of the exponential family (Fischer, 1974), conditional maximum likelihood (CML) estimation is another possibility. In this sense, the Rasch model is unique in IRT, and therefore we will discuss CML in some detail.

*Estimating the Rasch Model*

Let $\xi = \exp(\theta)$ and $\varepsilon_j = \exp(-\delta_j)$; then (16.25) becomes

$$P_j(\xi) = \frac{\xi\varepsilon_j}{1+\xi\varepsilon_j}. \tag{16.26}$$

Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ and $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_J)$. Also, let $x_{v+} = \sum_{j=1}^{J} x_{vj}$ and $x_{+j} = \sum_{v=1}^{N} x_{vj}$. Using this notation, the likelihood in (16.17) can be written as

$$\begin{aligned} L(\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}|\boldsymbol{\xi}, \boldsymbol{\epsilon}) &= \prod_{v=1}^{N}\prod_{j=1}^{J} P_j(\xi)^{x_{vj}}[1 - P_j(\xi)]^{1-x_{vj}} \\ &= \frac{\prod_{v=1}^{N} \xi_v^{x_{v+}} \prod_{j=1}^{J} \varepsilon_j^{x_{+j}}}{\prod_{v=1}^{N}\prod_{j=1}^{J}(1+\xi_v\varepsilon_j)}. \end{aligned} \tag{16.27}$$

Note that in (16.27) the marginals of the data matrix $\mathbf{X}_{N\times J}$ are sufficient statistics for estimation of the latent parameters of the model. CML proceeds as follows.

Let the person marginals of $\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}$ be collected in $\mathbf{x}_{N+} = (x_{1+}, \ldots, x_{N+})$. Consider the

probability of the data conditional on these person marginals $\mathbf{x}_{N+}$,

$$P(\mathbf{X}_{N\times J}=\mathbf{x}_{N\times J}|\mathbf{x}_{N+};\boldsymbol{\xi},\boldsymbol{\epsilon})=\frac{P(\mathbf{X}_{N\times J}=\mathbf{x}_{N\times J}|\boldsymbol{\xi},\boldsymbol{\epsilon})}{P(\mathbf{x}_{N+}|\boldsymbol{\xi},\boldsymbol{\epsilon})}. \tag{16.28}$$

Standard texts on the Rasch model (e.g., Fischer, 1974) explain in much detail that this equation can be shown to depend only on the item parameters $\boldsymbol{\epsilon}$ and the sufficient statistics for $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$ but not on $\boldsymbol{\xi}$. The resulting equation is then taken as a so-called conditional likelihood and solved for $\boldsymbol{\epsilon}$; this yields CML estimates for $\boldsymbol{\epsilon}$. These CML estimates are consistent and close to being maximally efficient (Eggen, 2000). In practice, the CML estimates of $\boldsymbol{\epsilon}$ are used to estimate $\boldsymbol{\xi}$ by means of ML (for details, see Fischer, 1974; Hoijtink & Boomsma, 1995). Fisher's information functions for items and tests are found by deriving the typical expressions for the Rasch model for $I_j(\theta)$ (16.22) and $I(\theta)$ (16.23).

CML thus enables us to estimate one set of parameters independently of the other set. Statistically, this is known as parameter separability. At the theoretical level, the possibility of making statements about items irrespective of the person distribution and, reversely, about persons irrespective of the difficulty level of the test is known as specific objectivity. Fischer (1974) considers specific objectivity crucial for measurement, but this point of view also has met with criticism (e.g., Borsboom, 2005). At the practical level, parameter separation is considered convenient for equation of scales, constructing item banks, and adaptive testing; these topics are discussed later.

### *Fitting the Rasch Model*

Goodness-of-fit tests for the Rasch model have been summarized by Glas and Verhelst (1995). Here, we mention the asymptotic $\chi^2$ tests, $R_1$ and $R_2$. The $R_1$ statistic tests the null hypothesis that the $J$ IRFs are parallel logistic curves as in (16.25), and $R_2$ tests whether WLI (16.2) holds under the Rasch model for all $\frac{1}{2}J(J-1)$ item pairs simultaneously. Rejection of parallel logistic curves for all items simultaneously could be indicative of different slopes between IRFs, and the approximate standard normal statistic called $U_j$ (Molenaar, 1983) may be used to test whether observed IRFs are steeper (say, $U_j < -1.645$) or flatter (say, $U_j > 1.645$) than expected under the Rasch model. Rejection of WLI (indicated by a significant $R_2$) may be taken as evidence of multidimensionality.

### *Computer Software*

The program RSP (Glas & Ellis, 1993) was used in our data example. Item parameters were estimated using CML, and person parameters using ML assuming item parameter estimates to be the true values. Fit of the Rasch model to $J$ items was assessed using statistics $R_1$ and $R_2$ and fit to individual items using statistic $U_j$.

## The Partial Credit Model

Masters's (1982) partial credit model (PCM) is an extension of the Rasch model to polytomous item scores and is often used for practical data analysis. For ordered, polytomous item scores, $0,\ldots,m$, the PCM models $m$ adjacent score pairs, $(0,1),\ldots,(m-1,m)$, as separate Rasch models:

$$\begin{aligned}P(X_j=x_j|\theta;X_j=x_j-1\vee X_j=x_j)&\\ =\frac{\exp(\theta-\delta_{jx_j})}{1+\exp(\theta-\delta_{jx_j})},&\qquad x_j=1,\ldots,m.\end{aligned} \tag{16.29}$$

As in the Rasch model, parameter $\delta_{jx_j}$ locates this response function on the $\theta$ scale, and for $\theta=\delta_{jx_j}$, the probabilities of having an item score of either $x_j-1$ or $x_j$ both equal 0.5. Combining the $m-1$ conditional probabilities in (16.29) yields the PCM

$$P(X_j=x_j|\theta)=\frac{\exp\left[\sum_{s=1}^{x_j}(\theta-\delta_{js})\right]}{\sum_{q=0}^{m}\exp\left[\sum_{s=1}^{q}(\theta-\delta_{js})\right]}. \tag{16.30}$$

Note that $x_j=0$ creates a problem in the numerator; hence, one chooses $\sum_{s=1}^{0}(\theta-\delta_{js})\equiv 0$, which results in $\sum_{x_j=0}^{m}P(X_j=x_j|\theta)\equiv 1$. This choice also defines $P(X_j=0|\theta)$ to be decreasing, which is seen as a desirable property.

Masters (1982) used data matrix $\mathbf{X}_{N\times J}$ with elements $x_{vj} = 0,\ldots,m$, and decomposed item scores $x_{vj}$ into $m$ binary scores $x_{vjs}$ $(s = 1,\ldots,m)$, with $x_{vjs} = 1$ if $x_{vj} = s$, and $x_{vjs} = 0$ otherwise. The marginal person total scores $x_{v+} = \sum_{j=1}^{J} x_{vj}$, with $v = 1,\ldots,N$, are sufficient statistics for estimating $\boldsymbol{\theta}_N$. The counts for each separate score on item $j$, $x_{+js} = \sum_{v=1}^{N} x_{vjs}$, with $s = 1,\ldots,m$, are the $m$ sufficient statistics for each of the parameters $\delta_{js}$, with $s = 1,\ldots,m$. Following a two-stage approach, CML is used for estimating the item parameters and ML for estimating the person parameters assuming that the item parameter estimates are the true values. Goodness-of-fit assessment is directed primarily at evaluating response functions, but the investigation of Assumption LI has met with considerable numerical problems. The program OPLM (Verhelst, Glas, & Verstralen, 1994) can be used for estimating and fitting the PCM.

## Comparing IRT Models

The 3PLM, the 2PLM and its generalization, the GRM, and the Rasch model and its generalization, the PCM, define response curves by means of parametric functions—here, logistic functions. Hence, these are parametric IRT models. Within the classes of models for dichotomous item scores, different models take different sets of item parameters into account. Hence, they provide descriptions of the data at different levels of complexity, each allowing for interesting explanations of the responses provided by the respondents. The GRM and the PCM provide models for different response probabilities but define similar item parameters.

Unlike parametric IRT models, the MHM imposes order restrictions on response functions, thus leaving them free to vary as long as Assumption M is satisfied. This is a nonparametric IRT model (e.g., Junker, 2001; Sijtsma & Meijer, 2007; Stout, 2002). In general, nonparametric models are based on weaker assumptions than parametric models. This is true within the set of models discussed here but not between any pair of nonparametric and parametric IRT models conceivable (e.g., Hemker et al., 1997).

The MHM is more general than the parametric models discussed here. For dichotomous-item models, in the nested sequence MHM-3PLM-2PLM-1PLM, each next model is a special case of the previous model. For polytomous-item models, Hemker et al. (1997) have shown not only that the GRM and the PCM are both special cases of the MHM but also that they do not imply one another and, even stronger, that they cannot be true simultaneously: There is no set of response functions in one model that can be transformed into another set that also satisfies the other model. However, differences between these models are often so small that in real-data analysis it may be difficult to distinguish the fit of one model from that of the other.

Because of their generality, nonparametric models have proven to be excellent starting points for deriving properties of IRT models in general (e.g., Ellis & Van den Wollenberg, 1993; Hemker et al., 1997; Holland & Rosenbaum, 1986; Junker, 1991, 1993; Stout, 2002). For example, Ellis and Van den Wollenberg (1993) showed that IRT models in general are true for subpopulations that have the same $\theta$ but not for individual respondents. This implies that for a particular $\theta$ value, say $\theta_d$, a response probability like $P(X_j = 1|\theta_d) = 0.7$ (dichotomous scoring) means that 70% of the respondents having the same $\theta_d$ provide a 1 score and 30% a 0 score, whereas the same individual is assumed to provide the same item score across independent replications. This is the random sampling interpretation of response probabilities (Holland, 1990). This interpretation contradicts general notions about human behavior, which assume that individuals show variation in response to the same item. This would imply the stochastic subject interpretation of response probability (Holland, 1990): $P(X_j = 1|\theta_d) = 0.7$ now means that respondent $v$ produces a 1 score in 70% and a 0 score in 30% of the random draws from his or her personal distribution of scores on item $j$. This has led Borsboom (2005) to argue that models for individual performance should be based on locally independent, repeated measurements, but he also noted that such repetitions usually are not available. This is a challenging conclusion that will need more attention in future developments.

Also interesting is the stochastic ordering result in (16.4). SOL holds for all dichotomous-item (M, LI, $D = 1$) models, including logistic models and also the MHM, which allows for irregular IRFs that are flat in some regions of $\theta$ and jagged elsewhere. SOL also holds for the PCM but not for the GRM and the polytomous-item MHM (Hemker et al., 1997).

Finally, for sets of either dichotomous or polytomous items that subsume under relaxed versions of each of the assumptions in (M, LI, $D = 1$), if $J \to \infty$, then the total score $X_+$ is a consistent ordinal estimator of $\theta$ (Junker, 1991). Thus, in nearly each IRT model, there is an intimate relationship between ordering according to the observable $X_+$ and the latent $\theta$, even in models that do not imply the SOL property. This suggests that, in general, little harm is done if the intuitively sensible total score $X_+$ is used for ordering persons (Van der Ark, 2005) under nearly any model that either assumes (M, LI, $D = 1$) or even violates these assumptions in controlled ways.

Due to the complexity of many test data sets, IRT models, either nonparametric or parametric, will not readily fit at the first attempt unless the data set is supported by sound empirical research that is based on well-articulated substantive theory. However, in most research, this is more the exception than the rule. Thus, IRT models are often rejected, which marks the beginning of multiple, complicated rounds of data analysis, in which several likely possibilities—leaving out items, trying subdivisions of the item set, fitting other models—are tried and overfitting is a realistic danger. Nevertheless, such data exploration may yield an acceptable result that, although it is different from what one had in mind at the outset, may provide a better understanding of what caused the model misfit. On the other hand, it rarely happens that a researcher starts an item analysis without at least a hunch or, better, an idea about the structure of his or her test. So rather than adopting a purely exploratory attitude, in practice, researchers often will look for a confirmation of their expectations and not just take any outcome for granted.

Even though nonparametric models are often considered exploratory and parametric models confirmatory data tools, in our opinion both approaches basically are used in the same way when analyzing complex test data. Nonparametric models may be a little more "open minded" because they use item selection procedures such as Mokken's (1971) and because they estimate the full response function, thus allowing many peculiarities of the data to become visible (Ramsay, 1991). Thus, in this sense, they are exploratory methods that let the data "speak for themselves." However, in those cases in which the researcher expects his or her item set to be (M, LI, $D = 1$), Assumptions LI and $D = 1$ can be evaluated using methods proposed by Stout et al. (1996) (not discussed here), Assumption M can be tested using the regression of an item score on the rest score (16.13), and measurement quality can be assessed using the $H$ and $H_j$ coefficients. Thus, the same methods that were considered exploratory tools when the researcher did not have a strong belief about his or her data have become confirmatory tools for testing his or her hypothesis about the test.

More than nonparametric models, probably due to their orientation toward statistical model testing, parametric models are often considered null hypotheses that are evaluated by means of formal statistical tests for the fit of the model to the data. First, statistical tests are used to find out whether a particular assumption of the model fits the data for all $J$ items simultaneously. For example, the 1PLM may be evaluated by means of the $R_1$ statistic, which assesses whether $J$ IRFs are parallel logistic curves, and the $R_2$ statistic, which assesses WLI for all $\frac{1}{2}J(J-1)$ item pairs. Second, because models often are found not to fit the data for the whole $J$-item test, one starts searching for items that could be deleted such that the model fits the data of the remaining item subset, one tries to find a subdivision of the item set in dimensionally distinct item clusters, or one uses other models to explain the data structure. This may involve several rounds of statistical testing of particular aspects of the model on (parts of) the data.

Thus, both parametric and nonparametric IRT data analyses often proceed in an exploratory rather than confirmatory manner, and as with most analyses of complex, highly multivariate data, the nature of the process depends much on whether one has strong hypotheses about one's measurement instruments or not.

**Table 16.1** MHM Analysis Results—"Scale Analysis": $P_j$ Values and $H_j$ Values for Total Test (15 Items); "Dimensionality Assessment": $H_j$ Values for Several Lower-Bound $c$ Values; and Total $H$ Values (Last Row)

| | *Scale Analysis* | | *Dimensionality Assessment* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $j$ | $P_j$ | $H_j$ | $c = 0.30$ | $c = 0.35$ | $c = 0.40$ | | $c = 0.45$ | | |
| 1 | 0.34 | 0.17 | us | us | us | us | us | us | us |
| 2 | 0.48 | 0.31 | 0.33 | us | us | us | us | us | us |
| 3 | 0.31 | 0.35 | 0.37 | 0.39 | 0.42 | — | 0.47 | — | — |
| 4 | 0.48 | 0.36 | 0.39 | 0.41 | — | 0.45 | us | us | us |
| 5 | 0.24 | 0.40 | 0.43 | 0.42 | — | 0.47 | — | 0.52 | — |
| 6 | 0.42 | 0.36 | 0.40 | 0.43 | 0.45 | — | us | us | us |
| 7 | 0.64 | 0.26 | us | us | us | us | us | us | us |
| 8 | 0.22 | 0.43 | 0.46 | 0.46 | 0.48 | — | 0.46 | — | — |
| 9 | 0.90 | 0.31 | 0.33 | 0.36 | 0.41 | — | — | — | 0.60 |
| 10 | 0.68 | 0.35 | 0.36 | 0.39 | 0.43 | — | 0.55 | — | — |
| 11 | 0.12 | 0.48 | 0.49 | 0.48 | 0.51 | — | 0.52 | — | — |
| 12 | 0.32 | 0.37 | 0.40 | 0.41 | — | 0.42 | — | — | 0.60 |
| 13 | 0.08 | 0.46 | 0.48 | 0.46 | 0.48 | — | — | 0.52 | — |
| 14 | 0.64 | 0.31 | 0.36 | 0.40 | us | us | us | us | us |
| 15 | 0.75 | 0.28 | 0.31 | us | — | 0.42 | — | 0.65 | — |
| Total $H$ | | 0.34 | 0.39 | 0.42 | 0.45 | 0.44 | 0.49 | 0.56 | 0.60 |

NOTE: "us" means the item was unscalable due to negative $H_{jk}$ with one of the selected items or because the $H_j$ value was smaller than lower-bound $c$.

## A Practical Data Example: Arithmetic of Proportions and Ratios

The nested sequence of dichotomous-item MHM, 3PLM, 2PLM, and 1PLM was used to analyze correct (score 1)/incorrect (score 0) scores from a 15-item arithmetic test. Dutch primary school students ($N = 612$) were asked to solve problems involving proportions and ratios. A typical constructed-response item is "If 10 oranges cost \$7.50, what do 3 oranges cost?" The MHM, the 3PLM, the 2PLM, and the 1PLM were fitted, in that order. This order of analysis shows neatly that as models impose more structure on the data, from the point of view of the model this may lead to simpler results and a test that is "pure" in terms of formal, psychometric properties, but from the point of view of the data this may lead to a loss of items and thus a loss of information on classifying individuals on the basis of their test scores.

### MHM Analysis

The proportions of correct answers ($P$-values, second column of Table 16.1) varied greatly. Item 13 was the most difficult item (smallest $P$) and Item 9 the easiest item (largest $P$). The third column shows the 15 item-scalability values (all $H_j$s significantly larger than 0; test results not tabulated). In an MHM analysis, $c = 0.3$ is considered the minimum for inclusion of items in a scale (Sijtsma & Molenaar, 2002, chap. 5). Because $H_j < 0.3$ for three items and because $H_j$ was small for several other items, the possibility of nonmonotonicities in the IRFs was evaluated next.

As an example, we discuss the IRF of Item 15 ($H_{15} = 0.28$). Its low $H_j$ value does not contradict the MHM but suggests that this item contributes little to an accurate person ordering. This suggestion would be supported by a violation of Assumption M. The discrete IRF estimate (Figure 16.1: left-hand panel, adapted from MSP) shows a significant decrease between two groups

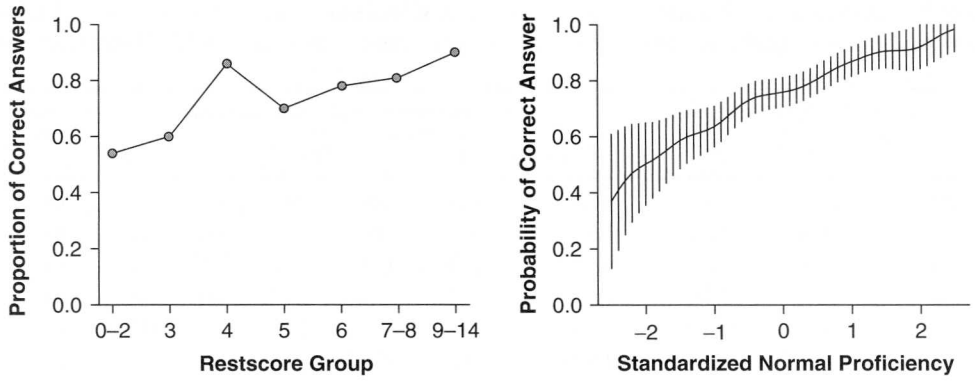


**Figure 16.1** IRF estimates of Item 15.

NOTE: (Left-hand side panel) Discrete estimate. (Right-hand side panel) Quasi-continuous estimate. Vertical bars indicate 95% confidence intervals.

based on rest scores (16.13) without Item 15, for that reason called restscore groups (two-tailed normal 5% test of $P[X_{15} = 1|R_{(-15)} = 4] = P[X_{15} = 1|R_{(-15)} = 5]$; $Z = 2.11$, $p = 0.035$). The continuous IRF estimate (right-hand panel, adapted from TestGraf98) is based on weighted averages of $P[X_{15} = 1|R_{(-15)}]$ across neighboring rest-score groups, but it does not pick up this violation. Based on this result and the large $p$-value of the normal test, we do not take this violation very seriously. Because none of the other 14 estimated IRFs showed significant violations, we conclude that the data support Assumption M for all items.

To assess dimensionality, items were clustered using $c = 0.30, 0.35, 0.40$, and 0.45. For $c = 0.30$, a 13-item cluster without Items 1 and 7 was found (note that now $H_{15} = 0.31$). Compared with the 15-item test, $H$ increased from 0.34 to 0.39. The IRFs estimated from the data without Items 1 and 7 did not show violations of Assumption M. Thus, the 13-item cluster satisfies the MHM and allows for sufficiently accurate measurement. A higher $c$ of 0.35 led to the additional rejection of Items 2 and 15. For $c = 0.40$, two clusters were found, and for $c = 0.45$, three clusters were found, while several other items proved unscalable. This sequence of outcomes—first, (nearly) all items are in the same cluster, and later, the cluster is split into smaller clusters while other items are unscalable—is taken as evidence of unidimensionality ($D = 1$) (Hemker et al., 1995). This conclusion was corroborated by inspection of the item content, which was highly similar both for items that were in the same cluster and for items that were in different clusters.

## 3PLM and 2PLM Analysis

Based on the MHM analysis, Items 1 and 7 were removed from the test, and then BILOG-MG (Zimowski et al., 1996; default settings were used) was used to first fit the 3PLM to the data and then the 2PLM. The $\gamma$ estimates ranged from 0.011 (SE = 0.008) to 0.102 (SE = 0.068); none was significantly larger than 0. A likelihood ratio test that compared the fit of the 2PLM and the 3PLM resulted in $\chi^2_{(df=13)} = 17.98$ ($p = 0.16$); hence, the fit of the models could not be distinguished, which confirmed that the $\gamma$ parameters could be dropped. In addition, for the 2PLM, the RMSD item fit statistics did not suggest misfit for any of the 13 items. Table 16.2 shows the $\alpha$ and $\delta$ estimates. Figure 16.2 (solid curve) shows the standard error for $\hat{\theta}$ (i.e., $I(\theta)^{-1/2}$; see (16.23)) based on all 13 items. The highest measurement precision was obtained for $0.75 < \theta < 1.00$. Thus, the test seems to measure the most accurate at the higher region of the scale.

**Table 16.2** Estimated Item Parameters and Item Fit Statistics for the 2PLM and the Rasch Model

| | *2PLM / MML Estimation* | | | | | *Rasch /CML Estimation* | | |
|---|---|---|---|---|---|---|---|---|
| $j$ | $\hat{\alpha}_j$ | $(SE)$ | $\hat{\delta}_j$ | $(SE)$ | $RMSD$ | $\hat{\delta}_j$ | $(SE)$ | $U_j$ |
| 2 | 0.73 | (0.09) | 0.11 | (0.09) | 1.20 | — | — | — |
| 3 | 0.90 | (0.10) | 0.73 | (0.09) | 0.60 | −0.38 | (0.10) | 0.36 |
| 4 | 0.92 | (0.10) | 0.07 | (0.08) | 1.45 | −1.43 | (0.10) | 0.49 |
| 5 | 1.08 | (0.14) | 0.97 | (0.10) | 0.76 | 0.17 | (0.11) | −0.41 |
| 6 | 1.00 | (0.14) | 0.30 | (0.08) | 0.66 | −1.04 | (0.10) | 0.04 |
| 8 | 1.24 | (0.16) | 0.99 | (0.09) | 0.42 | 0.27 | (0.11) | −0.31 |
| 9 | 0.52 | (0.09) | −2.80 | (0.44) | 0.94 | — | — | — |
| 10 | 0.69 | (0.08) | −0.83 | (0.11) | 0.64 | — | — | — |
| 11 | 1.11 | (0.17) | 1.56 | (0.15) | 1.37 | 1.11 | (0.13) | −0.53 |
| 12 | 1.01 | (0.12) | 0.67 | (0.09) | 0.81 | −0.44 | (0.10) | −0.15 |
| 13 | 1.01 | (0.15) | 1.95 | (0.19) | 0.62 | 1.74 | (0.15) | 0.39 |
| 14 | 0.72 | (0.09) | −0.60 | (0.10) | 0.94 | — | — | — |
| 15 | 0.52 | (0.08) | −1.46 | (0.21) | 0.75 | — | — | — |

NOTE: Item difficulties under MML and CML have been estimated using different norming of latent variable scale.

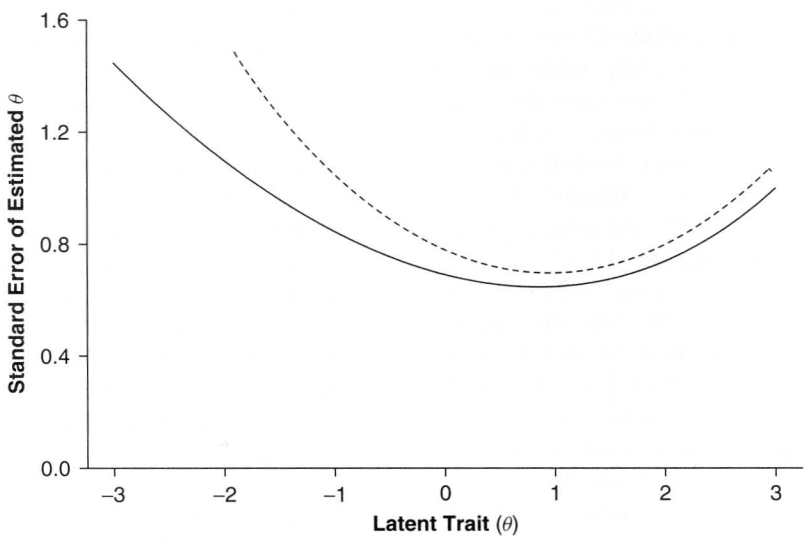


**Figure 16.2** Standard error of $\hat{\theta}$ based on 13 2PLM items (solid curve) and 8 Rasch items (dashed curve).

### Rasch Analysis

The different slope parameters found in the 2PLM analysis suggest that the Rasch model will not fit the data for all 13 items. Indeed, using RSP (Glas & Ellis, 1993) resulted in significant misfit ($R_1 = 64.20$, $df = 36$, $p = 0.003$; $R_2 = 128.07$, $df = 72$, $p = 0.000$). Next, a subset of 8 items with nearly the same IRF slopes was selected (based on the $\hat{\alpha}$s; the resulting item subset can be found in Table 16.2). For this subset, RSP analysis supported the hypothesis of equal slopes ($R_1 = 29.44$, $df = 21$, $p = 0.10$). This was further corroborated by the standard normal $U_j$ values ($|U_j| < 1.645$ for all $j$). In addition, support was obtained for Assumption LI ($R_2 = 36.62$, $df = 24$, $p = 0.05$). Figure 16.2 (dashed curve) shows the standard error (i.e., $I(\theta)^{-1/2}$) for $\hat{\theta}$ based on the 8 items. Eight Rasch items selected from 13 2PLM items necessarily provide less statistical information for the ML estimation of $\theta$, but the loss of precision was small because the 5 excluded items had relatively flat IRFs.

### Summary of Data Analysis

The MHM analysis of the 15 items showed that the $H_j$s were relatively small and the IRFs were monotone. A dimensionality analysis using varying lower-bound $c$ values suggested that 13 items with $H_j \geqslant 0.3$ together measured one latent trait. Together these results suggest that the IRFs had relatively weak, positive slopes and that the 13 selected items contributed modestly to accurate person ordering using $X_+$. The 3PLM analysis supported the conclusion that each of the IRF lower asymptotes was 0. The 2PLM fitted the data for the 13 items. A further selection of 8 items with approximately equal slopes led to a fitting Rasch model, but given that the items contributed modestly to person ordering, one may seriously wonder whether one is prepared to sacrifice 5 items to have a fitting Rasch model. Also, note that item selection was data driven and that this provides less compelling evidence for rejecting so many items than substantive reasons would.

## DISCUSSION

In this chapter, we have introduced IRT as a family of related models for measurement. We have concentrated on the analysis of data from a single test or questionnaire, because they represent the majority of IRT applications in most areas of science that use tests and questionnaires for measurement. This chapter has emphasized the estimation and fit evaluation of IRT models for single-test data, but several other analyses such as the following are possible:

- Differential item functioning (DIF) is aimed at checking whether an IRF or ISRF of an item is the same in different groups from the population of interest, such as boys and girls and different ethnic groups. If the response function of item $j$ is different, the item is said to exhibit DIF. DIF is often taken as a sign that in one group the item measures abilities or skills that are irrelevant for item performance in the other group. An example is an arithmetic test that also requires elementary language skills and a low-level group that varies considerably with respect to these language skills so that language skills level affects item performance differentially, whereas the individuals in the other group all have a skills level high enough not to affect item performance differentially.

- Person-fit analysis is aimed at identifying respondents who produce patterns of item scores, $\mathbf{x}$, that are atypical for the group to which they belong or relative to the IRT model that was fitted to the test data for this group. An example is students who have guessed excessively for correct answers in educational tests and thus produced patterns of 1s and 0s that are unrelated to item difficulty.

- Cognitive skills diagnosis is aimed at modeling the skills necessary to complete a task successfully or the cognitive process that underlies the response to a cognitive task. Models may formulate linear restrictions on item parameters to formalize skill contributions, assume multidimensional

latent variables to formalize a more complex ability structure, or assume a multiplicative structure for noncompensatory response processes. A fitting model provides information on skill deficiencies that require additional training or on the strategies used by children to solve complex cognitive problems. This may provide information on the developmental phase in which they are. This kind of information contributes to a better understanding of what a test measures—that is, to its validity.

Large-scale educational testing done by large testing agencies, such as Educational Testing Service (Princeton, NJ) and CITO National Institute for Educational Measurement (Arnhem, The Netherlands), uses possibilities offered by IRT such as the following:

- *Equating:* Items from different tests measuring the same ability or skill are displayed on a common scale, and students who have taken these different tests are comparable with one another on this scale and also with cutoff scores used for decision making.
- *Item banking:* This refers to the composition of a large set—hundreds—of items that measure the same ability or skill or sets of abilities and skills in a particular domain that is of interest for the evaluation of educational goals. An item bank contains the psychometric properties of all the items and also their position on a common, calibrated scale that is obtained by equating a large number of tests. In addition, an item bank contains information on item content, frequency of item use in tests, dates when they have been used, and so on, and together with the psychometric information on item difficulty and other item properties, this information can be used to assemble tests from the bank that simultaneously agree with a list of specifications deemed necessary for a particular application.
- *Adaptive testing:* This is aimed at presenting by computer to the examinee the smallest number of items that provides the most accurate estimate of his or her $\theta$ value in terms of $I(\theta)^{-1/2}$. The items are presented consecutively, and after each response, the estimation of $\theta$ is updated; the next item from the item bank to be presented is the one suited best to the examinee's $\hat{\theta}$ as we know it at that point in the testing process. This adaptive testing procedure stops if a formal criterion is satisfied—for example, if the standard error of $\hat{\theta}$ drops below a preset maximum value.

Equating, item banking, and adaptive testing require large-scale research and funding and are only feasible when the scale of the test application is such that the investments pay off. They are of more general interest because they make full use of the possibilities that parametric IRT models, in particular the 1PLM, the 2PLM, and the 3PLM, have to offer. Central is the possibility of equating scales, and the $\theta$ scale is a convenient tool for this.

## ACKNOWLEDGMENTS

The authors wish to express their gratitude to CITO National Institute for Educational Measurement for generously making the data available that we used for illustrating IRT model analysis.

## REFERENCES

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431–444.

Boomsma, A., Van Duijn, M. A. J., & Snijders, T. A. B. (2001). *Essays on item response theory.* New York: Springer-Verlag.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, *65*, 337–362.

Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models: A characterization of the homogeneous monotone IRT model. *Psychometrika*, *58*, 417–429.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern, Switzerland: Huber.

Glas, C. A. W., & Ellis, J. L. (1993). *User's manual RSP: Rasch scaling program*. Groningen, The Netherlands: iecProGAMMA.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer-Verlag.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383–392.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, *19*, 337–352.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–347.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 53–68). New York: Springer-Verlag.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*, 1523–1543.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, *56*, 255–278.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *Annals of Statistics*, *21*, 1359–1378.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. B. and M. A. J. Van Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 247–276). New York: Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, *24*, 65–81.

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*, 110–125.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken Scale: A Critical Discussion." *Applied Psychological Measurement, 10*, 279–285.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, *48*, 49–72.

Molenaar, I. W. (1997). Nonparametric models for polytomous items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer-Verlag.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows: User's manual*. Groningen, The Netherlands: iecProGAMMA.

Neyman, J., & Scott, E. L. (1948). Consistent estimation from partially consistent observations. *Econometrica*, *16*, 1–32.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 611–630.

Ramsay, J. O. (2000). *A program for the graphical analysis of multiple choice test and questionnaire data*. Montreal, Quebec, Canada: McGill University, Department of Psychology.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.

Stout, W. F., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*, 331–354.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Thissen, D., Chen, W.-H., & Bock, R. D. (2003). *MULTILOG (Version 7.0)*. Lincolnwood, IL: Scientific Software International. Computer software.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

Van der Ark, L. A. (2005). Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models. *Psychometrika, 70*, 283–304.

Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1994). *OPLM: Computer program and manual*. Arnhem, The Netherlands: CITO.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintainance for binary items*. Chicago: Scientific Software.